Selecting Informative Genes from Leukemia Gene Expression Data using a Hybrid Approach for Cancer Classification

Mohd Saberi Mohamad, Safaai Deris, Siti Zaiton Mohd Hashim

Laboratory of Artificial Intelligence and Bioinformatics,
Software Engineering Department, Faculty of Computer Science and Information Systems,
Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia

Abstract—**The development of microarray-based high-throughput gene profiling has led to the hope that this technology could provide an efficient and accurate means of diagnosing and classifying cancers. However, the large amount of data generated by microarrays requires effective selection of informative genes for cancer classification. Key issue that needs to be addressed is a selection of small number of informative genes that contribute to a disease from the thousands of genes measured on microarrays. This work deals with finding the small subset of informative genes from gene expression microarray data which maximize the classification accuracy. We introduce an improved version of hybrid of genetic algorithm and support vector machine for genes selection and classification. We show that the classification accuracy of the proposed approach is superior to a number of current state-of-the-art methods of one widely used benchmark dataset. The informative genes from the best subset are validated and verified by comparing them with the biological results produced from biology and computer scientist researchers in order to explore the biological plausibility.**

Keywords—**Gene selection; classification; genetic algorithm; support vector machine; gene expression; microarray**

I. Introduction

Due to recent advances in biotechnology, gene expression can now be quantitatively monitored on a global scale. Gene expression data is created by a process known as microarray that yields a set of floating points and absolute values [1]. These values represent the activity level of each gene within an organism at a particular point of time and a typical dataset can often consist of thousands of genes [2]. Recent studies on molecular level classification of tissue have produced remarkable results and indicated that microarray gene expression could significantly aid in the development of efficient cancer diagnosis [3,4]. However, classification based on the microarray data confronts with more challenges. One of the major challenges is the overwhelming number of genes relative to the number of training samples in the datasets [2,4,5]. Most of the genes are not relevant to the distinction between different tissue types (classes) and introduce noise in the classification process, and thus potentially drown out the contribution of the relevant ones [4].

In the gene expression domain, the gene refers to the feature. Feature selection or gene selection can be defined as a task for selecting subsets of features that maximizes the classifier ability to classify samples [6,7]. Gene selection methods can be classified into two categories. If gene selection is carried out independently from the classification procedure, the method is said to follow a filter approach. Otherwise, it is said to follow a wrapper (hybrid) approach [2,4]. Most of previous works used filter approach for selecting genes since it was computationally more efficient than the hybrid approach [4,8]. The major drawback is that an optimal selection of genes may be independent from the inductive and representational biases of the learning algorithm. Therefore, hybrid approach usually provide better accuracy but computationally more expensive than filter approach [4,9].

This research finds a small subset of informative genes from gene expression data which maximize the classification accuracy in order to make a diagnosis far more likely to be widely deployed in a clinical. In this paper, we present an improved version of hybrid of genetic algorithm (GA) and support vector machine (SVM) classifier (GASVM-II) for genes selection and classification.

In Section 2, we describe a hybrid of GA and SVM classifier (GASVM) and introduce GASVM-II. In Section 3, we analyze the experimental results followed by conclusion in Section 4.

II. A Hybrid of Genetic Algorithm and Support Vector Machine Classifier (GASVM)

The overall hybrid method consists of two main components: GA [10] and SVM [11] classifier. The GA will select subsets of features and then the SVM classifier evaluates the subsets during a classification process. The result of the classification is used for the fitness value of GA. Fig. 1 shows flow chart of GASVM.

F. Ibrahim, N.A. Abu Osman, J. Usman and N.A. Kadri (Eds.): Biomed 06, IFMBE Proceedings 15, pp. 528-532, 2007
www.springerlink.com © Springer-Verlag Berlin Heidelberg 2007

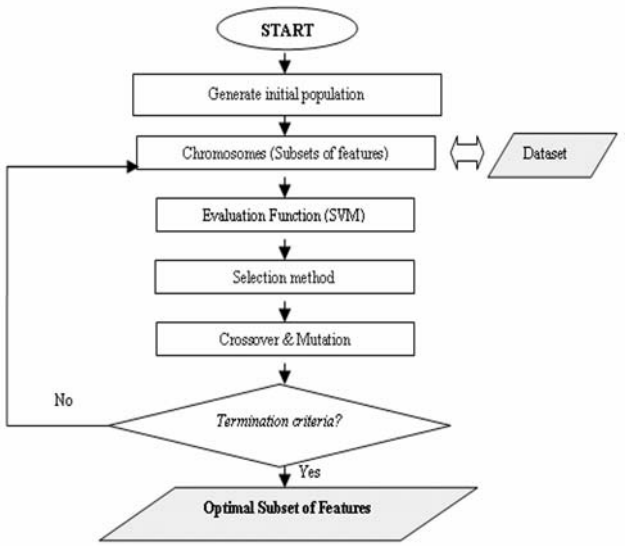

Fig. 1 A flow chart of hybrid of GA and SVM classifier (GASVM).

An individual represents a features subset (gene subset). The representation of chromosome (individual) used in GASVM appears in structural form as described in the previous works [2,9] and shown in Fig. 2.

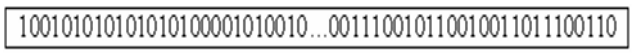

Fig. 2 A representation of chromosome in GASVM.

Let n be the total number of features available for representing the data to be classified. Hence, the chromosome is represented by binary vector of dimension n. If a bit is 1, it means that the corresponding feature is selected. A value of 0 indicates that the corresponding feature is not selected. The number of feature subsets based on the chromosome representation is calculated by using the following equation [10].

$$n_C = 2^n \tag{1}$$

where n_C is the number of feature subsets, whereas n is the number of features. A fitness function of each individual is determined by evaluating the SVM using a training set. Hence, this research has used a fitness function containing classification accuracy as mentioned below.

$$fitness(x) = accuracy(x) \tag{2}$$

where *accuracy(x)* is the *leave one out cross validation* (LOOCV) accuracy of the classifier with the features subset selection represented by x.

GA is used to maximize the fitness value in order to find the optimal features subset which has achieved the highest LOOCV accuracy. Finally, it produced the optimal subset of training set. The optimal subset from training set is used to construct SVM. Therefore, it is used to test the performance of built SVM.

A. An Improved Version of GASVM (GASVM-II)

Since the data used in this work is high dimensional data, the conventional approaches are hard to be applied. Hence, we proposed an improved chromosome representation in order to overcome the limitation. We have modified the representation of chromosome in GASVM for selecting subset of features suitable to gene expression data. The modified GASVM is called GASVM-II. This idea is based on reducing the number of feature subsets from the Equation (1) by fixing the number of selected features leading to this equation.

$$^{n}C_x = \frac{n!}{x!(n-x)!} \tag{3}$$

where $^{n}C_x$ is the total number of subsets of selected features x from the total of features n.

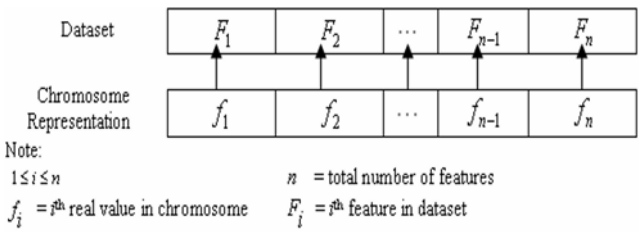

Fig. 3 An improved chromosome representation in GASVM-II

Fig. 3 shows an improved chromosome representation in GASVM-II that based on Equation (3). It includes the real value f_i in the chromosome which indicates a selected feature is the i^{th} feature among total features in dataset. For example, if $f_i = 10$, then the GASVM-II will select the 10^{th} feature in dataset in order to group it in related subset of features. The numbers of the real value f_i are equal to the numbers of selected features before evaluation process. This structure is not much affected by the total number of features and is able to represent chromosome in relatively small size. Its length can vary according to the size of the total number of features n and the number of selected features. The length of the chromosome is the same in size for each chromosome.

F. Ibrahim, N.A. Abu Osman, J. Usman and N.A. Kadri (Eds.): Biomed 06, IFMBE Proceedings 15, pp. 528-532, 2007
www.springerlink.com © Springer-Verlag Berlin Heidelberg 2007

The improved chromosome representation is developed to support more outstanding properties in genes selection for the cancer classification. The properties are as follows:

- Reducing the number of gene subsets.
- Supporting the high dimensional data.

III. Experimental results

Several set of experiments were conducted to compare the results of the SVM, GASVM and GASVM-II. We used one gene expression dataset, i.e., leukemia cancer dataset [12]. A LOOCV procedure is employed on training data and accuracy test measurement on testing data to measure classification accuracy [9,12].

A. Results Analysis and Discussions

The experiments of GASVM-II were conducted by using 10, 20, 30, 40, 50 and 60 genes in order to choose the best subset of genes among them.

Table 1 displays all subsets of genes can achieve high LOOCV accuracy due to high correlation between samples in training set. However, results of Leukemia dataset on accuracy test are not uniform. The datasets properties, i.e., thousand of genes with less than hundred of samples in the training sets can be possibly cause the overfitting which learning a decision surface that performs well on the training data but bad on testing data. Furthermore, most of the genes are not relevant to the distinction between different tissue types (classes) and introduce noise in the classification process, and thus potentially drowning out the contribution of the relevant ones.

Table 1. Classification accuracies for different gene subsets in Leukemia Cancer dataset using GASVM-II method.

Number of Selected Gene	Accuracy for Leukemia (%)	
	LOOCV	Test
10	100	79.4118
20	100	76.4706
30	100	94.1177
40	100	97.0588
50	100	94.1177
60	100	88.2353

Note:
The best result (subset of genes) shown in shaded cells

The selection of 40 genes from Leukemia Cancer dataset has achieved the best result at 100% using LOOCV procedure while 97.0588% using accuracy test measurement. Hence, this subset will be chosen as the best subset.

Table 2. Benchmark of GASVM-II and of previous methods on Leukemia Cancer dataset.

Method / Reference	Number of Selected Genes	Accuracy (%)	
		LOOCV	Test
GASVM-II	40	100	97.0588
GASVM	3568	94.7368	85.2941
SVM	7129	94.7368	85.2941
ART-NN [13]	10	100	97.0588
LD [14]	50	100	97.0588
MN [15]	10	100	90.0
SVM [16]	49	100	100
GAWV [9]	29	94.7368	88.2353
WV [12]	50	94.7368	85.2941

Note:
Methods in boldface were experimented in this research. The best results shown in shaded cells.

GASVM : A Hybrid of GA and SVM
GASVM-II : Proposed approach
SVM : Support vector machine classifier
WV : Weight voting classifier
GAWV : A Hybrid of GA and WV
ART-NN : Adaptive resonance theory neural network
LD : Logistic discriminant
MN : Modular neural network

Based on the LOOCV and the accuracy test in Table 2, it was noted that GASVM-II performance was equal to methods produced by [13], and [14]. However, the ART-NN method proposed by [13] is the best method because it produced acceptable result with smaller number of genes (10 genes) than other methods. Despite [16] achieved 100% accuracy by using 49 selected genes, but the result is not significant in this comparison because it rejected 4 samples when the confidence level procedure was introduced. Thus, the remaining of test samples only has 30 samples. The GASVM-II and several previous methods attained 100% accuracy using LOOCV procedure, but the previous researches cannot attain the same results when using testing accuracy manner [13,14,15]. This is due to overfitting the data during the training phase when learning a decision surface in the classifiers performed well on the training data but not for testing data. GASVM-II can classify 33 out of 34 test samples correctly. Among the sample, sample 66AML was consistently misclassified as AML. This AML sample was also misclassified by first original work [12] and other previous analyses [13,14].

All the previous works except [9] used filter approach for gene selection procedure. The filter approach is generally computationally more efficient than the hybrid approach. However, it was unable to avoid the noise and overfitting of the data because it is independent on classifier and depends on probabilistic distance measures, probabilistic dependence measures or interclass distance measures. Furthermore, the learning algorithm that has been used to construct the classifier was bias. Hence, the gene selection methods based on the filter approach caused the methods to perform poorly on classification of the datasets. [9] applied the hybrid ap-

F. Ibrahim, N.A. Abu Osman, J. Usman and N.A. Kadri (Eds.): Biomed 06, IFMBE Proceedings 15, pp. 528-532, 2007
www.springerlink.com © Springer-Verlag Berlin Heidelberg 2007

proach using GAWV. However, they required recurring experiment of the hybrid method to achieve an optimal subset. Moreover, the result is still less than others because this method was used the chromosome representation which is only supporting the data ranged from small to medium features.

As a result, when using the SVM classifier experimented in this research, the whole genes can contribute negative impact on classification performance because most of the genes in the data have many noises. GASVM method performs poorly because the chromosome representation was unable to fix the selected genes and impossible to search all feature spaces and evaluate all possible gene subsets. The GASVM is unable to evaluate all subsets due to huge number of subsets.

GASVM-II is able to avoid the noise problems because hybrid approach performs dependent on the classifier. The GASVM-II performs well in the experiment because it can fix the number of selected genes during gene selection and classification tasks. Hence, the GASVM-II reduces the complexity of search space and successfully evaluated all possible subsets of genes. It is shown that the selection of a small subset of informative genes using the GASVM-II can lead to significant improvement in classification accuracy for higher dimension data problems, i.e., gene expression data.

B. Biological Plausibility for Informative Genes in Leukemia Cancer Dataset

Biological plausibility is one of the criteria for causality in *epidemiology* [17]. It is prominent in all aspects of health risk assessments. A major goal of diagnostic research is to develop diagnostic procedures based on the least possible genes to detect diseases [2,9].

The best subset of 40 selected genes from Leukemia Cancer dataset was evaluated as the identical biological significant. These selected genes were evaluated by comparing them with the results produced from biologist and computer scientist researches.

Table 3. List of the same informative genes in Leukemia Cancer dataset produced by this research (GASVM-II) and previous works.

Previous Work	Gene Accession Number	Informative Gene Description
[18]	M13690	C1NH Complement component 1 inhibitor (angioedema, hereditary)
[3,13]	M55150	FAH Fumarylacetoacetate
[3,12,18]	M23197	CD33 antigen (differentiation antigen)
[3,18]	Y07604	Nucleoside-diphosphate kinase

Table 3 shows the lists of the similar informative genes of Leukemia Cancer dataset produced by GASVM-II and previous works. For instance, CD33 (M23197) were determined by [3,12,13], and [18]. CD33 is similarly a marker for AML, expressed in nearly all malignant *myeloblasts* [12].

From the Table 3, some of the informative genes produced by GASVM-II were validated and verified as the identical biology significance. Much time will be saved in finding and validating the genes by using the proposed approach than traditional *biopsy* procedure. A number of the genes identified by the GASVM-II in these experiments are already in use as clinical markers for cancer diagnosis. Some of the remaining genes may be excellent candidates for further clinical investigation. Thus, the GASVM-II has the ability to find out the informative genes to be used by medical and health sectors.

IV. Conclusion

We have investigated and solved the important issues of selecting a small subset of informative genes from thousands of gene measured on microarray that are inherently noisy. We have designed and developed the GASVM-II to select gene subsets for classification tasks.

Our experiments have empirically evaluated SVM, GASVM and GASVM-II using Leukemia Cancer dataset. The GASVM-II performs very well in most experiments. We are currently studying more on principle design of fitness using domain knowledge as well as mathematically well-founded tools.

Acknowledgments. This work was recognized by National Science Fellowship research program sponsored by Malaysian Ministry of Science, Technology and Environments (MOSTE).

V. References

1. Bannai, H., Inenaga, S., Shinohara, A., Takeda, M., Miyano, S.: Efficiently Finding Regulatory Elements Using Correlation with Gene Expression. J. Bioinfo. & Comput. Bio. 2 (2004) 273–288
2. Inza, I., Larranaga, P., Blanco, R., Correlaza, A.J.: Filter Versus Wrapper Gene Selection Approaches in DNA Microarray Domains. J. Art. Intel. Medic. 31 (2004) 91–103
3. Ben-Dor, A., Bruhn, L., Friedman, N., Schummer, I.M., Yakhini, Z.: Tissue Classification with Gene Expression Profiles. J. Comput. Bio. 7 (2000) 559–584
4. Wang, Y., Tetko, I.V., Hall, M.A., Frank, E., Facius, A., Mayer, K.F., Mewes, H.W. Gene Selection from Microarray Data for Cancer Classification – A Machine Learning Approach. J. Comput. Bio. & Chem. 29 (2005) 37–46.

F. Ibrahim, N.A. Abu Osman, J. Usman and N.A. Kadri (Eds.): Biomed 06, IFMBE Proceedings 15, pp. 528-532, 2007
www.springerlink.com © Springer-Verlag Berlin Heidelberg 2007

5. Soukup, M., Lee, J.K.: Developing Optimal Prediction Models for Cancer Classification using Gene Expression Data. J. Bioinfo. & Comput. Bio. 4 (2004) 681–694
6. Bins, J., Draper, B.A.: Feature Selection from Huge Feature Sets. Proc. Int. Conf. Comp. Vision. 2 (2001) 159–165
7. Zhang, P., Verma, B., Kumar, K.: Neural Vs Statistical Classifier in Conjunction with Genetic Algorithm Based Feature Selection. J. Patt. Recog Lett. 26 (2005) 909–919
8. Mohamad, M.S., Deris, S.: Feature Selection Method Using Genetic Algorithm for the Classification of Small and High Dimension Data. 1st Proc. Int. Symp. Info. Com. Tech. (2004) 13–16
9. Liu, J., Iba, H., Ishizuka, M.: Selecting Informative Genes with Parallel Genetic Algorithms in Tissue Classification. Genome Informatics. 12 (2001) 14–23
10. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. 3rd edn. Springer-Verlag, Berlin Heidelberg New York (1996)
11. Vapnik, V.: The Nature of Statistical Learning Theory. Springer-Verlag, New York (1995)
12. Golub, T.R., Slonim, D.K., Tomayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloofield, C.D., Lander, E. S.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Sci. 286 (1999) 531–537
13. Xu, R., Anagnostopoulos, G.C., Wunsch II, D.C.: Tissue Classification Through Analysis of Gene Expression Data Using a New Family of ART Architectures, Proc. Int. Joint Conf. Neu.Netw. (2002) 300–304
14. Nguyen, D.V., Rocke, D.M.: Tumor Classification by Partial Least Squares Using Microarray Gene Expression Data. Bioinformatics. 8 (2002) 39–50
15. Su, M., Basu, M., Toure, A.: Multi-Domain Gating Network for Classification of Cancer Cells Using Gene Expression Data. Proc. Int. Joint Conf. Neu. Netw. (2002) 286–289
16. Mukherjee, S.: Application of Statistical Learning Theory to DNA Microarray Analysis. PhD Thesis. Massachusetts Institute of Technology (2001)
17. Hill, A.B.: The Environment and Disease: Association or Causation. Proc. Royal Sci. Medic. (1965) 295–300
18. Krishnapuram,B., Carin, L., Hartemink, A.J.: Joint Classifier and Feature Optimization for Comprehensive Cancer Diagnosis Using Gene Expression Data. J. Comput. Bio. 11 (2004) 27–242

Address of the corresponding author:

Author: Mohd Saberi Mohamad
Institute: Universiti Teknologi Malaysia
Street: 81310 UTM Skudai
City: Johor Bharu, Johor
Country: Malaysia
Email: saberi@fsksm.utm.my

F. Ibrahim, N.A. Abu Osman, J. Usman and N.A. Kadri (Eds.): Biomed 06, IFMBE Proceedings 15, pp. 528-532, 2007
www.springerlink.com © Springer-Verlag Berlin Heidelberg 2007